Combination of Document Priors in Web Information Retrieval

Jie Peng and Iadh Ounis

Department of Computing Science, University of Glasgow, United Kingdom
{pj, ounis}@dcs.gla.ac.uk

Abstract. Query independent features (also called document priors), such as the number of incoming links to a document, its PageRank, or the length of its associated URL, have been explored to boost the retrieval effectiveness of Web Information Retrieval (IR) systems. The combination of such query independent features could further enhance the retrieval performance. However, most current combination approaches are based on heuristics, which ignore the possible dependence between the document priors. In this paper, we present a novel and robust method for combining document priors in a principled way. We use a conditional probability rule, which is derived from Kolmogorov's axioms. In particular, we investigate the retrieval performance attainable by our combination of priors method, in comparison to the use of single priors and a heuristic prior combination method. Furthermore, we examine when and how document priors should be combined.

1 Introduction

In Information Retrieval (IR), a document can have query-dependent and query-independent features. Query-dependent features relate to the characteristics of the document, which are specific to the queries and cannot be used before we receive the queries (e.g. the relevance of the document content to a given query). Query-independent features, also referred to as document priors, are features that do not depend on the queries. These document priors can be used to enhance the retrieval performance of a Web IR system, regardless of the query. For example, the number of incoming links to a document (Inlinks), its PageRank, or the length of its associated URL have been shown to be useful in some Web search tasks, such as the Homepage finding and Named Page finding tasks [2] [3].

The language modelling approach to IR provides an elegant framework to integrate single document priors into the retrieval process [3]. However, it is not clear how several document priors should be combined in a principled way. Indeed, most previous work considered either combining document prior probabilities in a heuristic way, usually assuming that document priors are independent from each other [3], or handtuning a linear combination of the priors [4]. Indeed, documents with a high PageRank score usually have a high number of incoming links, suggesting that the PageRank and Inlinks priors are often correlated. In addition, handtuning a linear combination of prior scores is heuristic, and not

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 732–736, 2007.

very practical in a realistic setting, where relevance judgements are not always available. In this paper, we present a novel and robust method for combining document priors in a principled way. We use a conditional probability rule, which is derived from Kolmogorov's axioms. The objective of the paper is two-fold: Firstly, we examine how effective our proposed method for the combination of priors is compared with the usually adopted heuristic approach. Secondly, we investigate whether the combination of document priors leads to a better retrieval performance compared to a baseline with and without the use of single priors. In particular, we examine when document priors should be combined.

2 Use of Single Priors in Language Modelling

In language modelling, the probability $P(D|Q)$ of a document D being generated by a query Q is estimated as follows [1]:

$$P(D|Q) = \frac{P(D) \cdot P(Q|D)}{P(Q)} \tag{1}$$

$P(D)$ is a document prior probability. $P(Q)$ can be ignored since it does not depend on the documents and, therefore, does not affect the ranking of documents. $P(Q|D)$ is given by [1]: $P(Q|D) = \sum_{i=1}^{n} \log(1 + \frac{\lambda \cdot tf(t_i,d)(\sum_t cf(t))}{(1-\lambda)cf(t_i)(\sum_t tf(t,d))})$, where λ is a constant given between 0 and 1. n is the number of query terms. $tf(t_i, d)$ is the term frequency of query term t_i in a document d; $\sum_t tf(t, d)$ is the length of document d, i.e. the number of tokens in the document; $cf(t_i)$ is the term frequency of query term t_i in the collection, and $\sum_t cf(t)$ is the total number of tokens in the collection.

In the above language modelling approach, the document prior $P(D)$ usually refers to a single document prior probability. This prior can be omitted from Equation (1), if all documents have a uniform prior. However, it is possible to consider multiple priors for each given document. In this case, it is important to combine the document prior probabilities in a principled way, taking into account the possible dependence between the considered document priors. In the next section, we propose a novel method for appropriately combining document priors.

3 Combination of Multiple Document Priors

For combining document priors, most of the current approaches either assume that the document priors are independent [3], or handtune a linear combination of the priors [4]. In the case the priors are assumed to be independent, the following formula is often used to combine two document priors p_1 and p_2:

$$P(D)_{p_1 \oplus p_2} = P(D)_{p_1} \cdot P(D)_{p_2} \tag{2}$$

where $P(D)_{p_1}$ is the document prior probability related to prior p_1; $P(D)_{p_2}$ is the document prior probability related to prior p_2; $P(D)_{p_1 \oplus p_2}$ is the document prior probability referring to the combination of both priors p_1 and p_2.

However, as mentioned in Section 1, the document priors are not necessarily independent. Therefore, we propose a different approach for combining the prior probabilities. We use a conditional probability rule that is based on Kolmogorov's axioms, given as follows:

$$P(D)_{p_1 \oplus p_2} = P(P(D)_{p_2}|P(D)_{p_1}) \cdot P(D)_{p_1} \tag{3}$$

where $P(D)_{p_1}$ is the document prior probability related to prior p_1, called the *base prior* probability; $P(P(D)_{p_2}|P(D)_{p_1})$ is the conditional probability related to prior p_2, given the prior p_1; $P(D)_{p_1 \oplus p_2}$ is the joint probability of both priors p_1 and p_2 occurring.

Note that the above conditional probability rule can be easily extended to more than two priors. $P(P(D)_{p_2}|P(D)_{p_1})$ can be estimated from a set of relevance judgements as follows: Firstly, we divide the prior probability $P(D)_{p_1}$ into several equal size bins on a logscale. Secondly, inside each bin, we divide the prior probability $P(D)_{p_2}$ into several equal size subset bins, again on a logscale. Finally, the conditional probability of $P(P(D)_{p_2}|P(D)_{p_1})$ in each subset bin is the number of target documents divided by the number of documents in that subset bin.

4 Experiments and Analysis

In this paper, we consider four well-established document priors, namely Page-Rank (PR), information-to-noise ratio (ITN) [5], document length (DL), and the document URL score [3]. We use the standard .GOV Web test collection, and its corresponding TREC 2003 and TREC 2004 Homepage and Named Page finding topic and relevance assessment sets. The official evaluation measure for both tasks is the Mean Reciprocal Rank (MRR).

Firstly, we assess the performance of each of the four single priors (see Table 1). Our baseline (BL) is a language modelling approach, where all documents have a uniform prior probability. From Table 1, we can see that, in general, the single document priors can improve the retrieval performance on the used tasks. The only exception is the ITN prior, which leads to a degradation of the retrieval performance in most cases.

Secondly, we investigate the combination of every pair of priors using our proposed combination approach, and compare it to the performance of the corresponding single priors. Note that the used base prior probability is important in Equation (3). From Table 1, we observe that several combinations of document priors lead to an enhanced MRR score, when we use an effective document prior as base. In particular, combining the best single priors usually leads to an enhanced retrieval performance, compared to their single use.

Finally, we compare our proposed method to a heuristic combination approach, where the priors are assumed to be independent. From Table 1, we observe that our combination way generally outperforms the heuristic method,

Table 1. MRR for the Named Page and Homepage tasks. We use $\lambda = 0.9$ in all experiments. The best retrieval performance is highlighted in bold, and the base prior probabilities are highlighted in italic. Runs statistically different from the best run (Wilcoxon Matched-Pairs Signed-Ranks Test, $p < 0.05$) are underlined. Note that for lack of space, only the most commonly used priors for each task are combined.

Named Page Finding			Homepage Finding		
	MRR			MRR	
	TREC 2003	TREC 2004		TREC 2003	TREC 2004
BL	0.4366	0.3533	BL	0.2363	0.1200
BL+PR	0.4539	0.3588	BL+PR	0.4339	0.3558
BL+DL	0.4546	0.4116	BL+URL	0.4738	0.3976
BL+ITN	0.4186	0.3583	BL+ITN	0.1980	0.0980
Our Proposed Method					
BL+*PR*+DL	0.3730	0.3117	BL+*PR*+URL	0.5247	0.4062
BL+*DL*+PR	0.4732	**0.4365**	BL+*URL*+PR	**0.5424**	**0.4446**
BL+*PR*+ITN	0.3755	0.3098	BL+*PR*+ITN	0.4059	0.3385
BL+*ITN*+PR	**0.4894**	0.4021	BL+*ITN*+PR	0.3889	0.3696
BL+*DL*+ITN	0.4787	0.4130	BL+*URL*+ITN	0.4729	0.4133
BL+*ITN*+DL	0.4377	0.3495	BL+*ITN*+URL	0.4615	0.3508
Priors Independence Assumption Method					
BL+PR +DL	0.4674	0.4065	BL+PR+URL	0.5409	0.4110
BL+PR+ITN	0.4815	0.3867	BL+PR+ITN	0.3526	0.3166
BL+DL+ITN	0.4232	0.3704	BL+URL+ITN	0.4551	0.3470

when the best of the two combined document priors is used as the base prior. The only exception is related to the ITN prior, when it is used as a base prior to combine with the PageRank or URL prior. This combination seems to work very well. Further investigation is required to understand the behaviour of ITN. Overall, our proposed technique can always outperform the heuristic method.

The above results are consistent across both used retrieval tasks. In addition, we observe that, excepting for the TREC 2003 Named Page finding task, using the two best single priors leads to the best overall MRR performance.

5 Conclusion

We have investigated the retrieval performance attainable with query-independent features, in the form of document prior probabilities on two Web search tasks, using a standard Web test collection. We showed that our proposed conditional combination method increases the retrieval performance over the respective single priors, when we use the two best-performing single priors. In addition, we observed that our technique can always outperform a heuristic method, which assumes the independence of priors.

References

1. Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis. (2001)
2. Kamps, J., Mishne, G., de Rijke, M.: Language Models for Searching in Web Corpora. In Proc. of TREC 2004, Gaithersburg, MD, (2004)
3. Kraaij, W., Westerveld,T., Hiemstra, D.: The Importance of Prior Probabilities for Entry Page Search. In Proc. of SIGIR 2002, Finland, (2002)
4. Metzler, D., Strohman, T., Zhou, Y., Croft, W.B.: Indri at TREC 2005: Terabyte Track. In Proc. of TREC 2005, Gaithersburg, MD, (2005)
5. Zhu X.L., Gauch, S.: Incorporating Quality Metrics in Centralized / Distributed Information Retrieval on the WWW. In Proc. of SIGIR 2000, Athens, (2000)